ERROR PREDICTION IN SPOKEN DIALOG: FROM SIGNAL-TO-NOISE RATIO TO SEMANTIC CONFIDENCE SCORES

Dilek Hakkani-Tür, Gokhan Tur, Giuseppe Riccardi *Hong Kook Kim*

AT&T Labs - Research
Florham Park, NJ 07932
{dtur,gtur,dsp3}@research.att.com

Gwangju Institute of Science and Technology
Gwangju, 500-712, Korea
hongkook@gist.ac.kr

ABSTRACT

Spoken dialog systems aim to interpret meanings of users' utterances and respond to them accordingly. The users' utterances are first recognized by an automatic speech recognizer (ASR) and the intents of the users are extracted by the spoken language understanding (SLU) unit. Both ASR and SLU are noisy and in general their noise statistics are *not* correlated. Our goal is to exploit the signal-noise information and ASR lattice-based and semantic confidence scores for SLU error prediction and prevention of these by rejecting erroneous utterances, or asking confirmation questions. In our experiments, we have shown up to 80% relative decrease in the error rate of the accepted utterances collected using the AT&T How May I Help You^{TM} Spoken Dialog System used for customer care.

1. INTRODUCTION

Spoken dialog systems aim to identify intents of humans, expressed in natural language, and take actions accordingly, to satisfy their requests [1]. In a natural spoken dialog system, typically, first the speaker's utterance is recognized using an automatic speech recognizer (ASR). Then, the intent of the speaker is identified from the recognized sequence, using a spoken language understanding (SLU) component. This step can be framed as a classification problem for call routing systems [1, 2, among others]. For example, if the user says "*I would like to know my account balance*", then the corresponding intent or semantic label (call-type) would be "*Request(Balance)*", and the action would be prompting the user's balance, after getting the account number, or transferring the user to the billing department.

For each utterance in the dialog, SLU returns a call-type associated with a confidence score. If the SLU confidence score is more than the confirmation threshold, the dialog manager takes the appropriate action as in the example above. If the intent is vague, the user is asked a clarification prompt by the dialog manager. If the SLU is not confident about the intent, the utterance is either simply rejected by re-prompting the user (if the confidence score is less than the rejection threshold) or a confirmation prompt is played (if the SLU confidence score is in between confirmation and rejection thresholds).

It is clear that SLU confidence score is critical for the spoken dialog management. On the other hand, relying solely on SLU confidence scores for determining the dialog strategy is suboptimal, because of several reasons. First of all, with spontaneous telephone speech, the typical word error rate (WER) for ASR output is around 25-30%; in other words, one in every three to four words is misrecognized [3, 4]. Misrecognizing a word may result in misunderstanding the whole utterance, even though all other words are correct, such as misrecognizing the word "*balance*" in the utterance above. Second, SLU confidence scores may depend on the estimated call-type and other utterance features such as the length of utterance in words, or contextual features, such as the previous prompt played.

In the literature, there is a large number of studies on ASR confidence score estimation using acoustic and linguistic features. Our aim is to detect SLU errors due to noisy ASR output or other reasons, and prevent further problems that may result due to these errors with the help of these confidence scores. In addition to the SLU confidence score, we compute stationary and time-varying quantity of noise, and ASR lattice-based confidence scores which are obtained from acoustic and language models. We have used various ways of combining these information sources, such as logistic regression and decision trees, with the aim of detecting the correctly interpreted utterances.

Combining multiple information sources for error detection in spoken dialog systems includes the following work: Langkilde *et al.* [5] used various features like the recognized utterances, their length, confidence scores of call-types, prompts, type of the prompts, etc. in order to find problematic dialogs. Our approach is very similar to this, however, instead of finding problematic dialogs, we search for problematic utterances on-line, and help the dialog manager to take the appropriate action *during* the dialog. An additional difference, is that, we also utilize ASR and signal-noise re-

0-7803-8874-7/05/$20.00 ©2005 IEEE

lated features when making the decision. Hazen *et al.* also exploited ASR confidence scores and other utterance related features, like the hypothesized number of words, to reject utterances [6]. However, they did not consider the SLU confidence scores or dialog level features. Moreover, they made a binary decision of rejection and acceptance, ignoring confirmations. Raymond *et al.* combined the confidence scores obtained from the acoustic model, language model, and semantic model for better spoken language understanding via re-scoring the semantic n-best hypotheses [7]. Our previous work focused on using word confidence scores to improve the decision of various classifiers [8], and improving the SLU confidence scores using logistic regression with utterance related features [9].

In the next section, we describe the computation of the confidence scores we have employed, and how we improve dialog strategies by combining these with other features. Then we present experimental results on AT&T Spoken Dialog System for customer care applications.

2. APPROACH

In spoken dialog systems, once the utterance is recognized, spoken language understanding examines each utterance, $\hat{w} = w_1, w_2, ..., w_n$, and assigns the utterance an intent (or a call-type), $\hat{c}(\hat{w})$, as well as a confidence score, $e(\hat{c})$, obtained from the semantic classifier. This score is used to guide the dialog strategies. If the intent is not vague and the score is higher than some threshold t_1 (that is, $e(\hat{c}) > t_1$), then the decision is accepted by the dialog manager, and the appropriate action is taken. If the score is lower than another threshold t_2 (that is, $e(\hat{c}) < t_2$), then the utterance is rejected, and the user is re-prompted. If the score is in between the two thresholds (that is, $t_2 \leq e(\hat{c}) \leq t_1$), then the user is asked a confirmation question to verify the estimated intent. These thresholds are selected to optimize the spoken dialog performance.

In order to be more robust to ASR errors, and improve acceptance, confirmation and rejection strategies during spoken dialog processing, we propose combining ASR lattice-based, and semantic confidence scores and stationary and time-varying quantity of noise. Note that it is straightforward to augment this set with dialog level features (such as turn number), utterance level features (such as utterance length in words), prosodic features, other semantic level features (such as the top scoring intent), etc. In this work we have only focused on the errors due to ASR.

2.1. Stationary and Time-Varying Quantity of Noise

The stationary quantity of noise in each utterance is measured as a signal-to-noise ratio (SNR), where signal and noise powers are estimated by detecting background noise. In this paper, the noise detection is through forced alignment, either by preserving the state segmentations during recognition, or by performing forced alignment with recognized transcriptions [10]. Let $I(n)$ be the identifier for the n^{th} analysis frame of a given utterance, where $I(n) = 1$ indicates that the n^{th} frame belongs to the speech interval, and $I(n) = 0$ indicates that the n^{th} frame belongs to the silent interval. If the number of frames for the utterance is L, the average log energy for the speech intervals, SP, is given by:

$$SP = \frac{1}{\sum_{n=1}^{L} I(n)} \sum_{n=1}^{L} E(n)|_{I(n)=1}$$

where $E(n)$ is the log energy of the n^{th} frame, which is defined as $10 \log_{10} \sum_{i=1}^{N} s^2(i)$ where N is the number of samples in a frame and $s(i)$ is a sample value. Similarly, the average log energy for the silent intervals, NP, is computed by:

$$NP = \frac{1}{L - \sum_{n=1}^{L} I(n)} \sum_{n=1}^{L} E(n)|_{I(n)=0}$$

As a result, the stationary SNR measurement for the utterance is estimated as the difference between the average log energies of speech and silent intervals: $SNR = SP - NP$.

The stationary SNR measurement does not reflect the local characteristics of environmental noise and in many cases such measurements can be very misleading. For example, although the utterance is with high SNR, the average SNR measurement would be low due to the highly non-stationary noise signal in the last part of that utterance. Tracking such non-stationarities is essential for providing a more accurate SNR measurement. A non-stationary measure to quantify the variation of SNR along an utterance, which is referred to as non-stationary SNR (NSNR). NSNR is defined as the standard deviation of noise power normalized by the average signal power, and computed by:

$$NSNR = (\frac{1}{L - \sum_{n=1}^{L} I(n)} \sum_{n=1}^{L} (SP - E(n))^2|_{I(n)=0} - SNR^2)^{\frac{1}{2}}$$

NSNR, which is measured in dB, becomes smaller when the average of the frame-dependent SNR, defined by $(SP - E(n))$, approaches the SNR measurement. This implies that smaller variations in the noise characteristics among different frames would result in low measurement of NSNR.

2.2. ASR Lattice-Based Scores

In order to capture the scores obtained from acoustic and language models, we compute word posterior probabilities for each word w_j, of each utterance from the lattice output of ASR, where $j = 1, ..., n$. We use these posterior probabilities as word confidence scores cs_j for each word w_j. We use the word confidence scores cs_j to assign an ASR score to the utterance, $e(\hat{w}) = f(cs_1, ..., cs_n)$ where f is the arithmetic mean function.

The algorithm for computing word confidence scores is based on the *pivot* alignment for strings in the word lattice. A detailed explanation of this algorithm and the comparison of its performance with other approaches is presented in [11].

2.3. Semantic Scores

In AT&T spoken dialog system, the goal of the SLU unit is understanding the intent of the user, which can be framed as a classification problem [1]. Given a set of examples $S = \{(W_1, c_1), ..., (W_m, c_m)\}$, the problem is to associate each instance $W_i \in X$ into a target label $c_i \in C$, where C is a finite set of call-types that are compiled automatically or semi-automatically from the data. However, it is often useful to associate some confidence score to each of the classes. In this work, we have employed a discriminative classifier, namely Boostexter [12]. This is an implementation of the AdaBoost algorithm, which iteratively learns simple weak base classifiers [13]. Friedman *et al.* have suggested a method for converting the output of AdaBoost to confidence scores using a logistic function [14]:

$$P(c = c_j|W_i) = \frac{1}{(1 + e^{-2\times f(c_j,W_i)})}$$

where $f(c_j, W_i)$ is the weighted average score of the base classifiers produced by AdaBoost for utterance W_i and call-type c_j. Then the SLU score for W_i is computed using the following equation: $e(\hat{c}) \approx \max_{c_j} P(c = c_j|W_i)$.

2.4. Combining Scores

We have converted the problem of estimating a better confidence score for each utterance into a classification problem, where we try to find a function to combine multiple features, and estimate a new score. For this purpose, we have employed logistic regression and decision trees and used the individual scores computed for the utterance as features.

3. EXPERIMENTS AND RESULTS

3.1. Data

We used AT&T How May I Help You spoken dialog system for automated customer care, in order to test our approach. There are 84 unique call-types in this application, and the test set call-type perplexity, computed using the prior call-type distribution estimated from the training data, is 32.64. We split the data into three sets: training set, development set, and test set. The first set is used for training the ASR language model and SLU model, which are then used to recognize and classify the other two sets. We used an off-the shelf acoustic model. The development set is used to estimate the parameters of the score combination function. We show our results on the test set. Some properties of these data sets are given in Table 1. SLU accuracy (*SLU Acc.*) is the percentage of utterances, whose top-scoring call-type is among the true call-types. The top-scoring call-type of an utterance, is the call-type that is given the highest score by the classifier. The true call-types are the call-types assigned by human labelers to each utterance.

In order to simulate the effect of this approach in a deployed application, we selected the test set from the latest days of data collection, therefore there is a mismatch in the performance of ASR and SLU on the two test sets. When we examined the data, we noticed a difference in the distribution of the call-types, due to changes in customer traffic.

	Training Set	**Dev. Set**	**Test Set**
No. of utterances	9,094	5,171	6,296
ASR Word Acc.	-	68.8%	70.2%
SLU Acc. (ASR)	-	65.65%	62.81%
SLU Acc. (Trans.)		75.22%	71.68%

Table 1. Some properties of the training, development and test data used in the experiments.

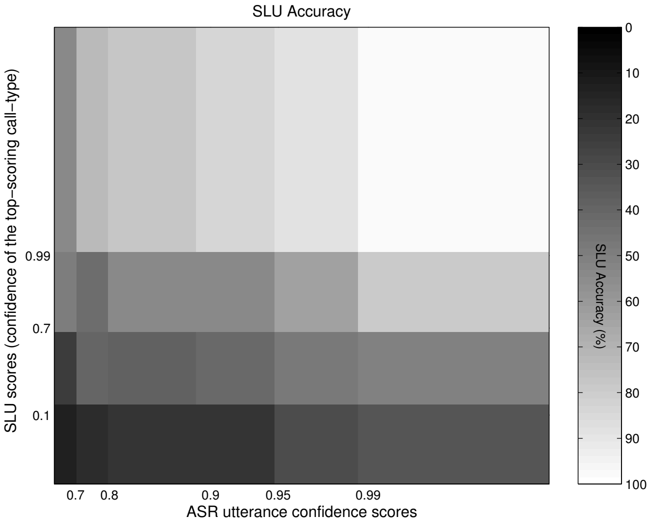

Fig. 1. The SLU accuracy for ASR and SLU confidence score bins. The color of each rectangle shows the SLU accuracy in that bin, and the size of each rectangle corresponds to the number of examples in that bin.

3.2. Evaluation Results

In order to check the feasibility of improving the accuracy of accepted utterances, we first plotted the SLU accuracy for various ASR lattice-based and semantic confidence score bins. Figure 1 shows the 4-dimensional plot for these bins, where x-axis is the lattice-based confidence score bin, y-axis is the semantic confidence score bin. The color of each rectangle, corresponding to these bins, shows the SLU accuracy in that bin, and the size of each rectangle is proportional to the number of examples in that bin. As can be seen from this plot, when the two scores are high, SLU accuracy is also high, when both of them are low, SLU accuracy is also low. But when the lattice-based confidence score is low, SLU accuracy is also low, even though SLU score is high. This figure proves that SLU score alone is not enough to determine the accuracy of the estimated intent.

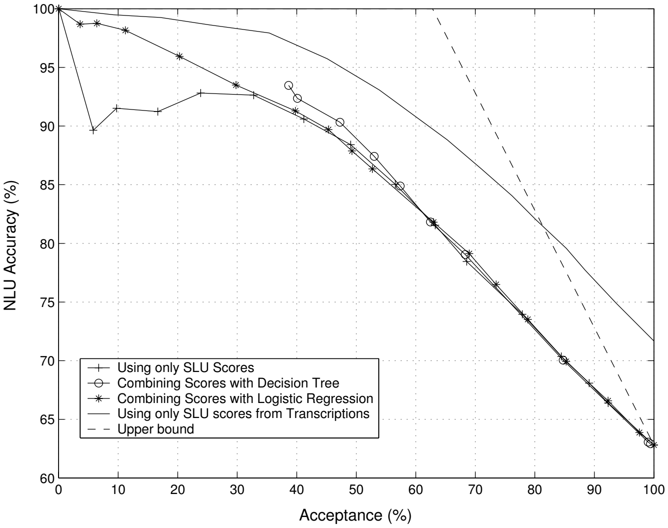

Fig. 2. The accuracy of accepted utterances, where the new score is used to accept utterances.

Using logistic regression, we have got the normalized coefficients presented in Table 2. This can be considered as another indicator of the importance of each of the features for the SLU error detection task.

Const	ASR	SLU	SNR	NSNR	Length
-4.957	4.516	2.531	-3.219	0.374	-0.341

Table 2. Normalized coefficients of the logistic regression.

Figure 2 presents the results of our experiments for combining multiple information sources. The x-axis is the percentage of the accepted utterances, and the y-axis is the percentage of utterances that are correctly classified. The baseline is using only the SLU scores for this purpose, and the upper bound is the cheating experiment, where we first reject all the erroneously classified utterances, by comparing them with their true call-types. As another upper bound, we have used the manual transcription of each utterance, and used only the SLU confidence score. Both methods for combining features with SLU confidence scores helped improving the accuracy of the accepted utterances. Due to the decision tree we have trained, the points until 40% acceptance rate is unaccessible. One impressive results is obtained using logistic regression for an acceptance rate of 5%. The rejection error rate has decreased 80% from around 10% to only 2%. Then this improvement vanishes as the acceptance rate increases as expected.

4. CONCLUSIONS

We have presented an approach for combining signal-noise, ASR and SLU confidence scores to detect understanding errors and prevention of these by rejecting erroneous utterances, or asking confirmation questions. We have evaluated this approach using a deployed AT&T Spoken Dialog System for a customer care application and have shown up to 80% relative decrease in the error rate of the accepted utterances. As a future work, we would like to extend the set of features to include prosodic, dialog and utterance level features.

5. REFERENCES

[1] A. L. Gorin, A. Abella, T. Alonso, G. Riccardi, and J. H. Wright, “Automated natural spoken dialog,” *IEEE Computer Magazine*, vol. 35, no. 4, pp. 51–56, April 2002.

[2] P. Natarajan, R. Prasad, B. Suhm, and D. McCarthy, “Speech enabled natural language call routing: BBN call director,” in *Proceedings of the ICSLP*, Denver, CO, September 2002.

[3] B. Kingsbury, L. Mangu, G. Saon, G. Zweig, S. Axelrod, V. Goel, K. Visweswariah, and M. Picheny, “Toward domain-independent conversational speech recognition,” in *Proceedings of the Eurospeech*, Geneva, Switzerland, September 2003.

[4] G. Riccardi and D. Hakkani-Tür, “Active and unsupervised learning for automatic speech recognition,” in *Proceedings of the Eurospeech*, Geneva, Switzerland, September 2003.

[5] I. Langkilde, M. A. Walker, J. Wright, A. Gorin, and D. Litman, “Automatic prediction of problematic human computer dialogues in how may i help you?,” in *Proceedings of IEEE ASRU Workshop*, 1999.

[6] T. J. Hazen, S. Seneff, and J. Polifroni, “Recognition confidence scoring and its use in speech understanding systems,” *Computer Speech and Language*, , no. 16, pp. 49–67, 2002.

[7] C. Raymond, F. Béchet, R. De Mori, and G. Damnati, “On the use of confidence for statistical decision in dialogue strategies,” in *Proceedings of the SigDial Workshop*, Boston, MA, May 2004.

[8] Gokhan Tur, Dilek Hakkani-Tür, and Giuseppe Riccardi, “Extending boosting for call classification using word confusion networks,” in *Proceedings of ICASSP*, Montreal, Canada, 2004.

[9] G. Tur, J. Wright, A. Gorin, G. Riccardi, and D. Hakkani-Tür, “Improving spoken language understanding using word confusion networks,” in *Proceedings of the ICSLP*, Denver, CO, September 2002.

[10] H. K. Kim and M. Rahim, “Why speech recognizers make errors? a robustness view,” in *Proceedings of the ICSLP*, Jeju Island, Korea, October 2004.

[11] D. Hakkani-Tür and G. Riccardi, “A general algorithm for word graph matrix decomposition,” in *Proceedings of the ICASSP*, Hong Kong, May 2003.

[12] R. E. Schapire and Y. Singer, “Boostexter: A boosting-based system for text categorization,” *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.

[13] R. E. Schapire, “The boosting approach to machine learning: An overview,” in *Proceedings of the MSRI Workshop on Nonlinear Estimation and Classification*, Berkeley, CA, March 2001.

[14] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: A statistical view of boosting,” *The Annals of Statistics*, vol. 38, no. 2, pp. 337–374, 2000.